Advanced Regression: 1c Random effects and hierarchical models (Part I)

Garyfallos Konstantinoudis

Epidemiology and Biostatistics, Imperial College London

21st February 2023

Motivation

- Structured data
- Individual-level and group-level

Fixed effect analysis

- Definition of fixed effects
- Fixed effects in R

Motivation

All methods presented so far assume that the observations are iid.

iid: Independent and identically distributed

- **Independent**: The observations are independent from each other

$$cor(x_i, x_{i'}) = 0 \text{ for all } i, i' \in 1, ..., n$$

- **Identically**: All observations have the same distribution. For example when assuming a Normal distribution they all have the same mean and variance.

PS: Exchangeability: Allows for dependence between observations and only states that future observations behave like past ones.

Motivation: How realistic is iid?

- Often our data contains structure depending on how our data was sampled.
  - Within K boroughs in London we select n participants ...
  - From K schools we sample n students ...
  - From K hospitals we select n patients ...
  - At K stores we sampled n costumers ...
- $k \in 1, ..., K$ group index

Grouping creates dependence

Observations within a group are likely to be more similar to each other than to observations from other groups.

Motivation: GP data

- We are interested in the relationship of cholesterol and age and how age impacts cholesterol.
- Sampling: We take measurements of patients from certain GPs.
  - Group-level: GPs $K = 12$ `table(data.chol[["doctor"]])`

```
 1  2  3  4  5  6  7  8  9 10 11 12
36 36 36 39 36 36 39 36 36 39 36 36
```

  - Individual-level: Patients $n = 441$

```
head(data.chol)
```

	chol	doctor	age	bmi	agedoc	sex
1	7.13	1	54	27.39	55	0
2	7.70	1	55	29.10	55	0
3	7.30	1	56	27.90	55	0
4	6.89	1	71	26.67	55	1
5	6.90	1	72	26.70	55	1
6	7.90	1	73	29.70	55	1

Pooled analysis

Linear model using all $i = 1, ..., n$ observations ignoring the grouping

$$y_i = \alpha_0 + \beta x_i + \epsilon_i$$

Assumptions

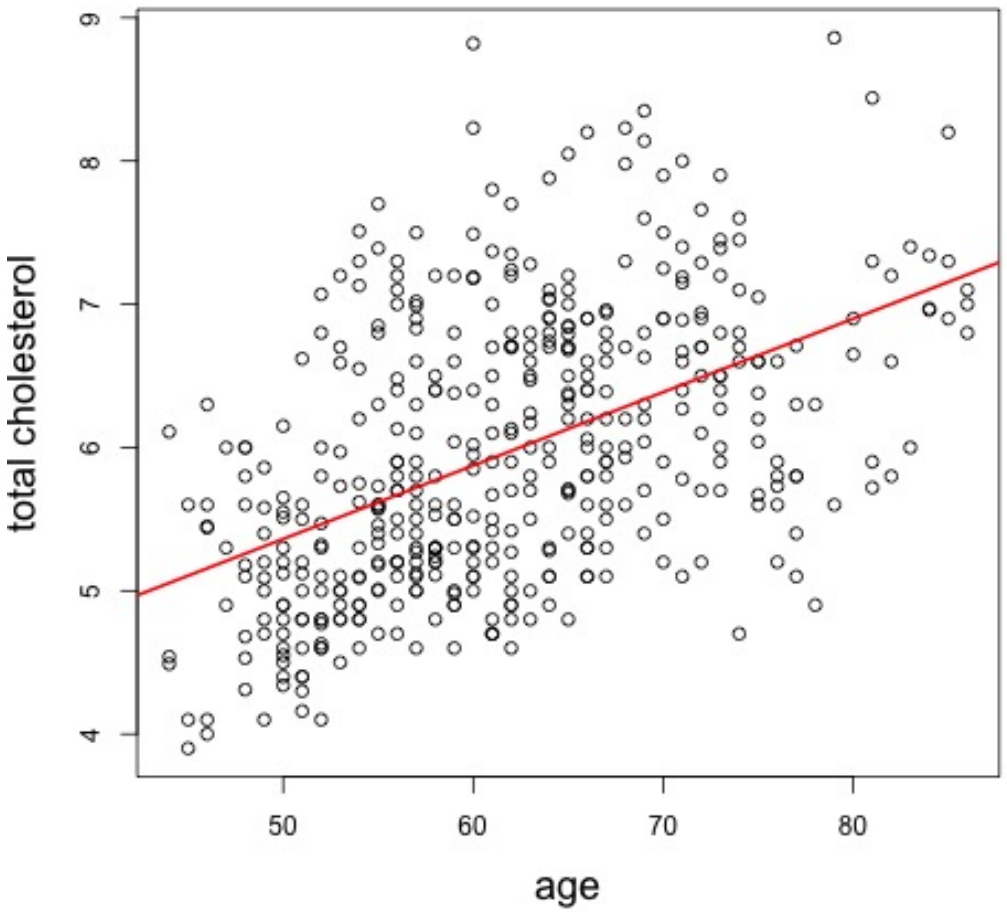
- All observations independent (incorrect).

Consequences

- Estimated errors on regression coefficients are too small.
- Overstate significance of association.

GP data: Pooled analysis

```
Pooled.Model = lm(chol ~ age, data=data.chol)
```



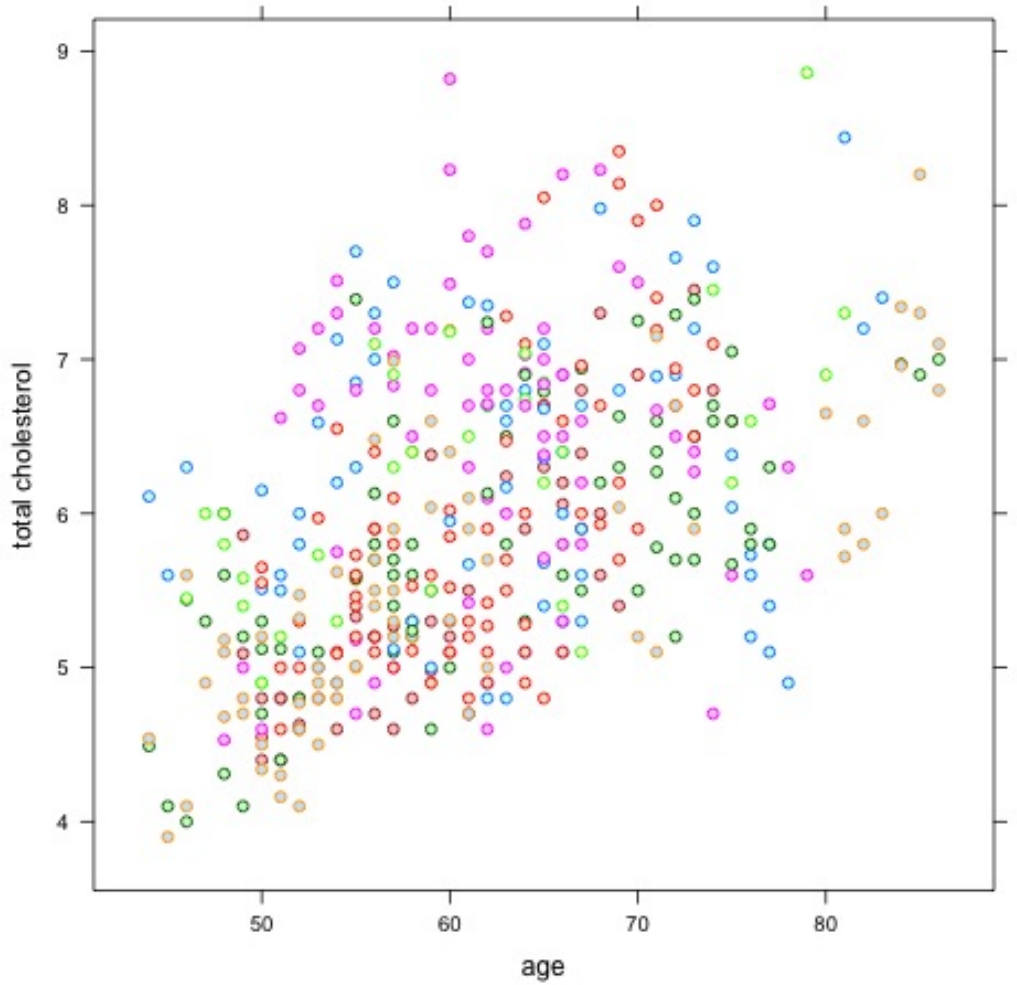
GP data: Pooled analysis

```
Pooled.Model = lm(chol ~ age, data=data.chol)
summary(Pooled.Model)
```

```
Call:
lm(formula = chol ~ age, data = data.chol)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8971 -0.6206 -0.1105  0.5693  2.9456

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.798691   0.268571   10.42   <2e-16 ***
age         0.051262   0.004301   11.92   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8362 on 439 degrees of freedom
Multiple R-squared:  0.2445,    Adjusted R-squared:  0.2428
F-statistic: 142.1 on 1 and 439 DF,  p-value: < 2.2e-16
```

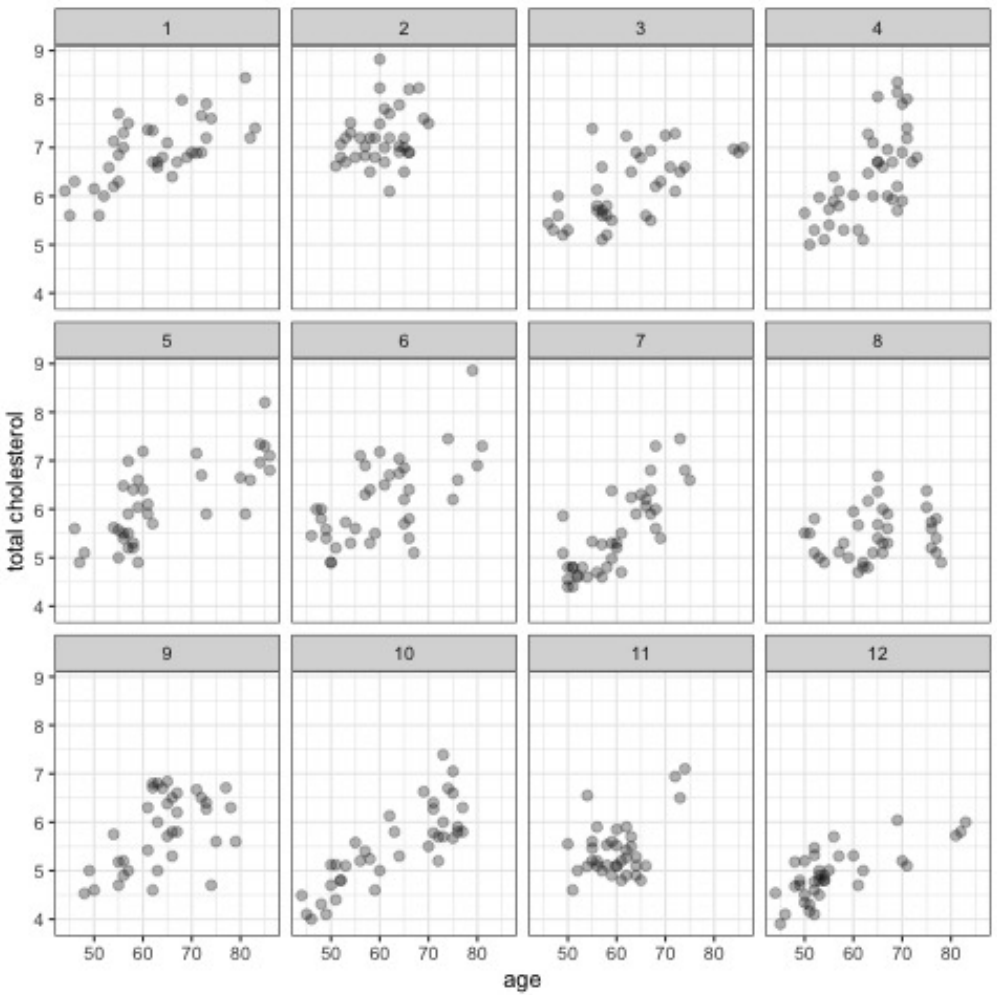
GP data: Pooled analysis

```
xyplot(chol~age, groups = doctor, data=data.chol,
pch = 21)
```



GP data: Pooled analysis

```
ggplot(data.chol, aes(x = age, y = chol, group =
doctor)) + facet_wrap(~doctor)
```



Ignoring dependence

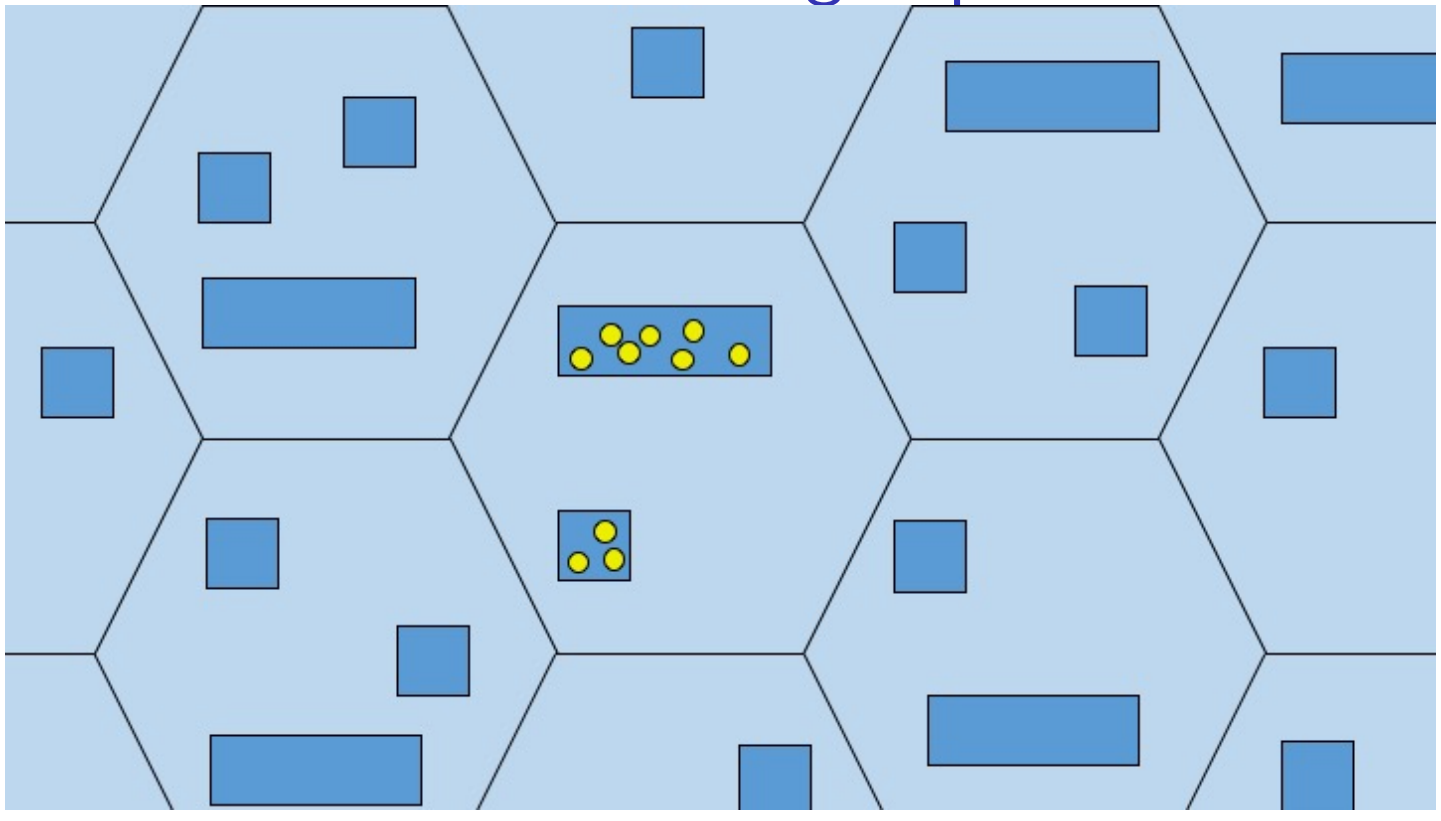
- standard errors too small
- $p-$values too small / confidence intervals too narrow
- over-estimate significance

Intuitively, there is less information in the data than an independent sample.

This has to be taken into account in our models:

1. Perform analysis for each group separately.
2. Calculate summary measures for each group and use standard analysis (Group-level analysis).
3. Fixed effects model to account for group structures.
4. Use random effects models that explicitly model the similarity of observations in a group.

Motivation: Individual-level and group-level



- Observations are grouped with grouping information known.
- Multi-level: Multiple levels of groupings, e.g. classrooms within schools within districts.
- Variables can be measured on the individual and group level.

1. Separate analysis

How to?

- Estimate separate regression coefficients for each group.

Assumptions

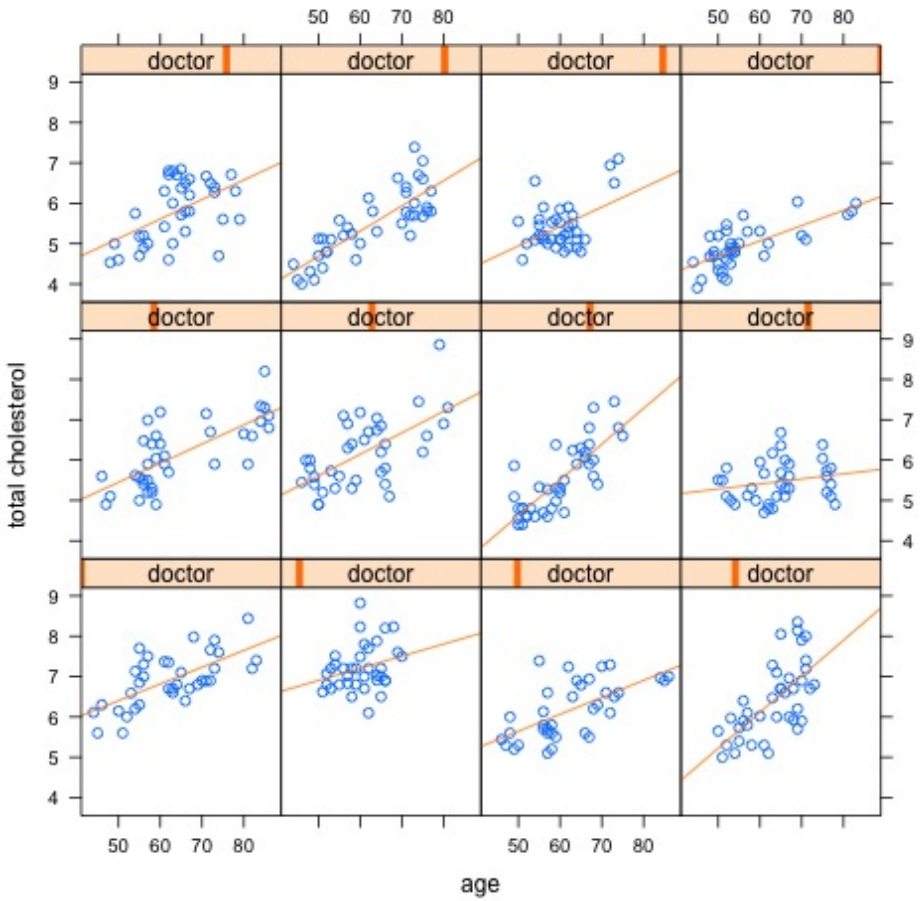
- Independence between groups.

Consequences

- This is a reasonable approach to exploratory analysis.
- If the number of individuals in each group is small, we will get imprecise estimates.
- Multiple testing is an issue.

GP data: Separate analysis

```
xyplot(chol ~ age | doctor, data=data.chol)
```



2. Group-level analysis

How to?

- Summarise outcome and predictors for each group k, e.g. using mean or median.

```
chol.group =
tapply(data.chol$chol,INDEX=data.chol$doctor,FUN=mean)
age.group =
tapply(data.chol$age,INDEX=data.chol$doctor,FUN=mean)
```

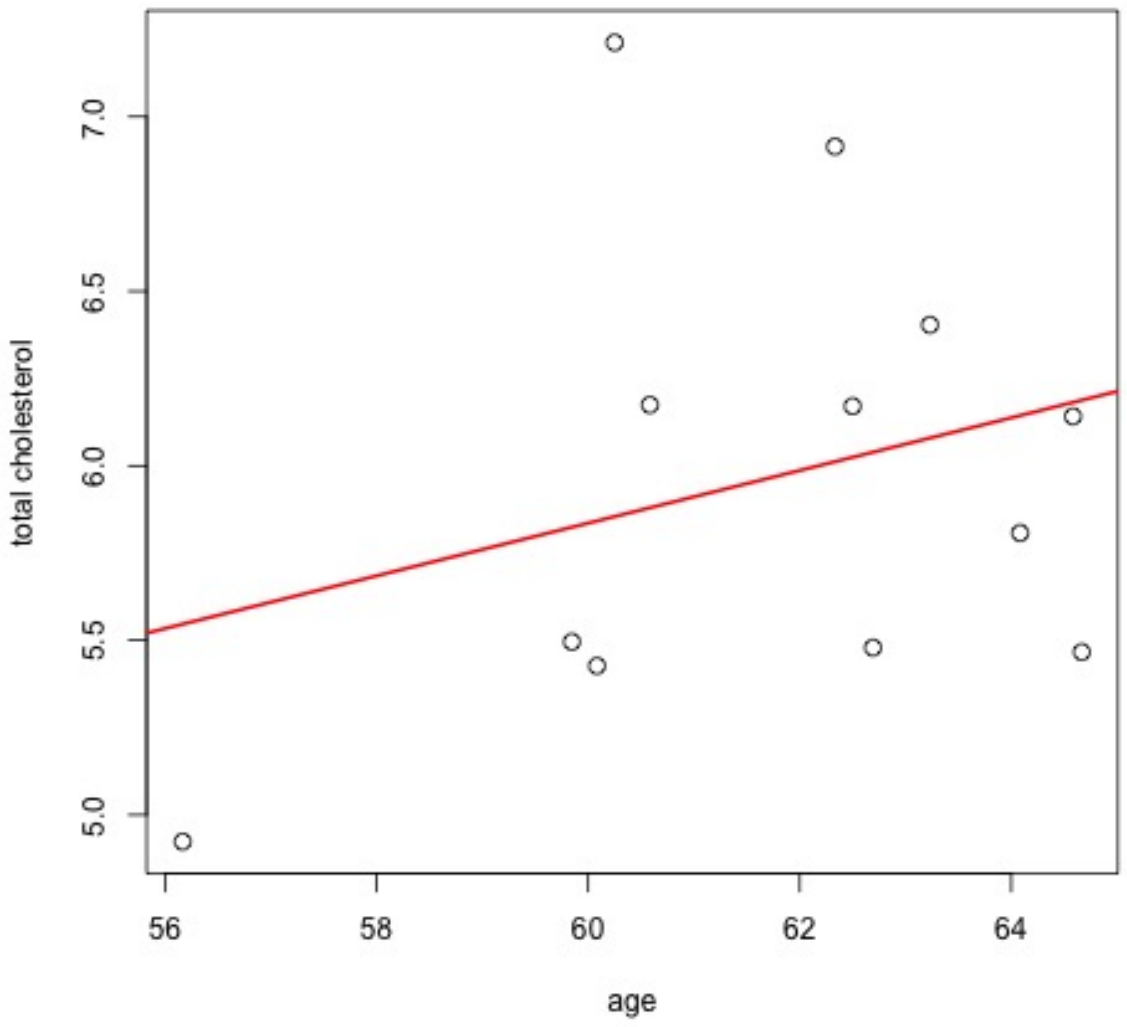
- Treat the group summaries as observations.

```
Group.Model = lm(chol.group ~ age.group)
summary(Group.Model)
```

Assumptions

- One regression line fit: Associations between outcome and predictors are the same for each group.
- Independence between groups.
- All groups are treated equal, irrespective of size.

GP data: Group level analysis



GP data: Group level analysis

```
Call:
lm(formula = chol.group ~ age.group)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7216 -0.4513 -0.1844  0.3020  1.3576

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.30687    5.05233   0.259    0.801
age.group    0.07548    0.08176   0.923    0.378

Residual standard error: 0.67 on 10 degrees of freedom
Multiple R-squared:  0.07854,   Adjusted R-squared:  -0.0136
F-statistic: 0.8524 on 1 and 10 DF,  p-value: 0.3776
```

Consequences

- This model lacks power as the number of data points used is the number of groups ($k < n$)
- Regression coefficients will be averaged over all groups $\rightarrow$ real within-group effects may be diluted.

Inverse variance weighted (IVW) meta-analysis

Each random variable is weighted in inverse proportion to its variance.

Assume we have independent observations y_k with variance σ_k. Then the IVW estimate is defined as

$$\hat{y}_{\text{IVW}} = \frac{\sum_{k=1}^{K} y_k/\sigma_k}{\sum_{k=1}^{K} 1/\sigma_k}$$

Weighted regression over groups

Assume y_k is a vector of group summaries, x_k is a $k \times p$ matrix of group summaries. Assume w is a diagonal matrix with $w[k,k] = \frac{1}{\sigma_k^2}$, then the weighted least squares estimate is defined as

$$\hat{\beta}_w = (x_k^t w x_k)^{-1} x_k^t w y_k$$

3. Fixed effects

Motivation:

- Keep the idea of modelling within groups: Allow associations to differ across groups.
- But now we model all the data (n observations) together: Maximise the power to detect associations.

Joint model with group-specific intercept

$$y_i = \alpha_k + \beta x_i + \epsilon_i$$

where α_k is a fixed effect.

- α_k captures the effect of unobserved group specific confounders.
- Residual errors $\epsilon_i, i \in 1, ..., n$ are assumed independent.

Fixed effects

How to?

- A fixed effects model is fit in the same way as the simple linear model including the group as a covariate.

Assumptions

- Information on α_k comes from observations in group k only.

Consequences

- By including group effects we have controlled for group characteristics.
- But introduced a large number of parameters (one for each group).
- May be a problem if there are few observations in some groups.

R: Fixed effects in lm()

- Fixed effects in `R` can be computed using the `lm()` model.
- Fixed effects are essentially categorical covariates (`as.factor()`).
- There are two different types of fixed effect:
  1. Group-specific intercept α_k

$$y_i = \alpha_k + \beta x_i + \epsilon_i$$

  2. Group-specific slope β_k

$$y_i = \alpha_0 + \beta_k x_i + \epsilon_i$$

R: Group-specific intercept in lm()

1. Group-specific intercept

$$y_i = \alpha_k + \beta x_i + \epsilon_i$$

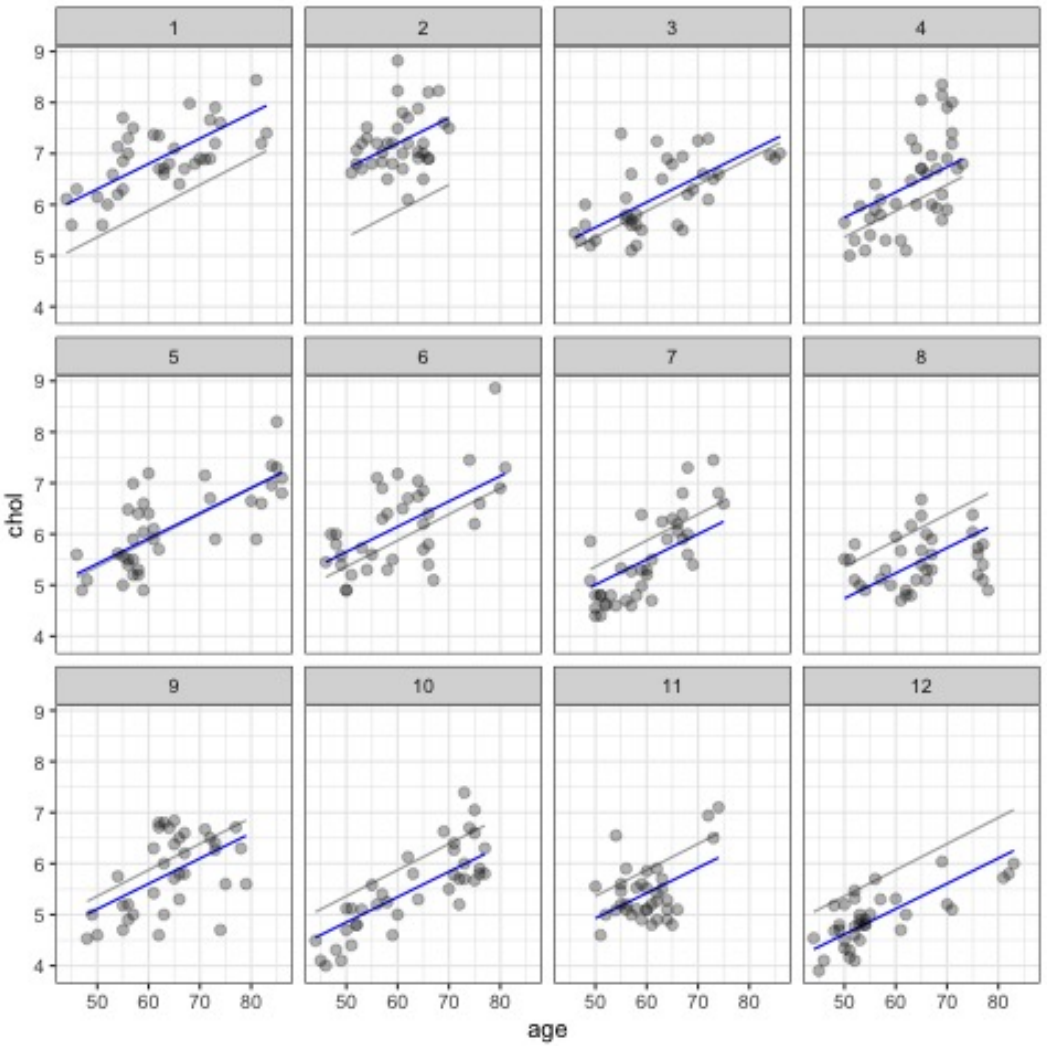
- Add the group variable as additional categorical (`as.factor()`) covariate.
- `Varying.Intercept.Model = lm(chol ~ age + as.factor(doctor), data=data.chol)`

R: Group-specific intercept in lm()

`summary(Varying.Intercept.Model)`

```
Call:
lm(formula = chol ~ age + as.factor(doctor), data = data.chol)

Residuals:
     Min       1Q   Median       3Q      Max
-1.59881 -0.40321 -0.08463  0.37929  1.77313

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.826236   0.213854  17.892  < 2e-16 ***
age                 0.049543   0.003065  16.164  < 2e-16 ***
as.factor(doctor)2  0.400993   0.136014   2.948  0.00337 **
as.factor(doctor)3 -0.752146   0.135865  -5.536 5.41e-08 ***
as.factor(doctor)4 -0.555317   0.133254  -4.167 3.73e-05 ***
as.factor(doctor)5 -0.884528   0.136039  -6.502 2.21e-10 ***
as.factor(doctor)6 -0.653299   0.135970  -4.805 2.15e-06 ***
as.factor(doctor)7 -1.295580   0.133444  -9.709  < 2e-16 ***
as.factor(doctor)8 -1.563657   0.136053 -11.493  < 2e-16 ***
as.factor(doctor)9 -1.193645   0.135970  -8.779  < 2e-16 ***
as.factor(doctor)10 -1.453255   0.133231 -10.908  < 2e-16 ***
as.factor(doctor)11 -1.376027   0.136039 -10.115  < 2e-16 ***
as.factor(doctor)12 -1.685593   0.137173 -12.288  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5764 on 428 degrees of freedom
Multiple R-squared:   0.65,     Adjusted R-squared:  0.6402
F-statistic: 66.24 on 12 and 428 DF,  p-value: < 2.2e-16
```

R: Group-specific intercept in lm()



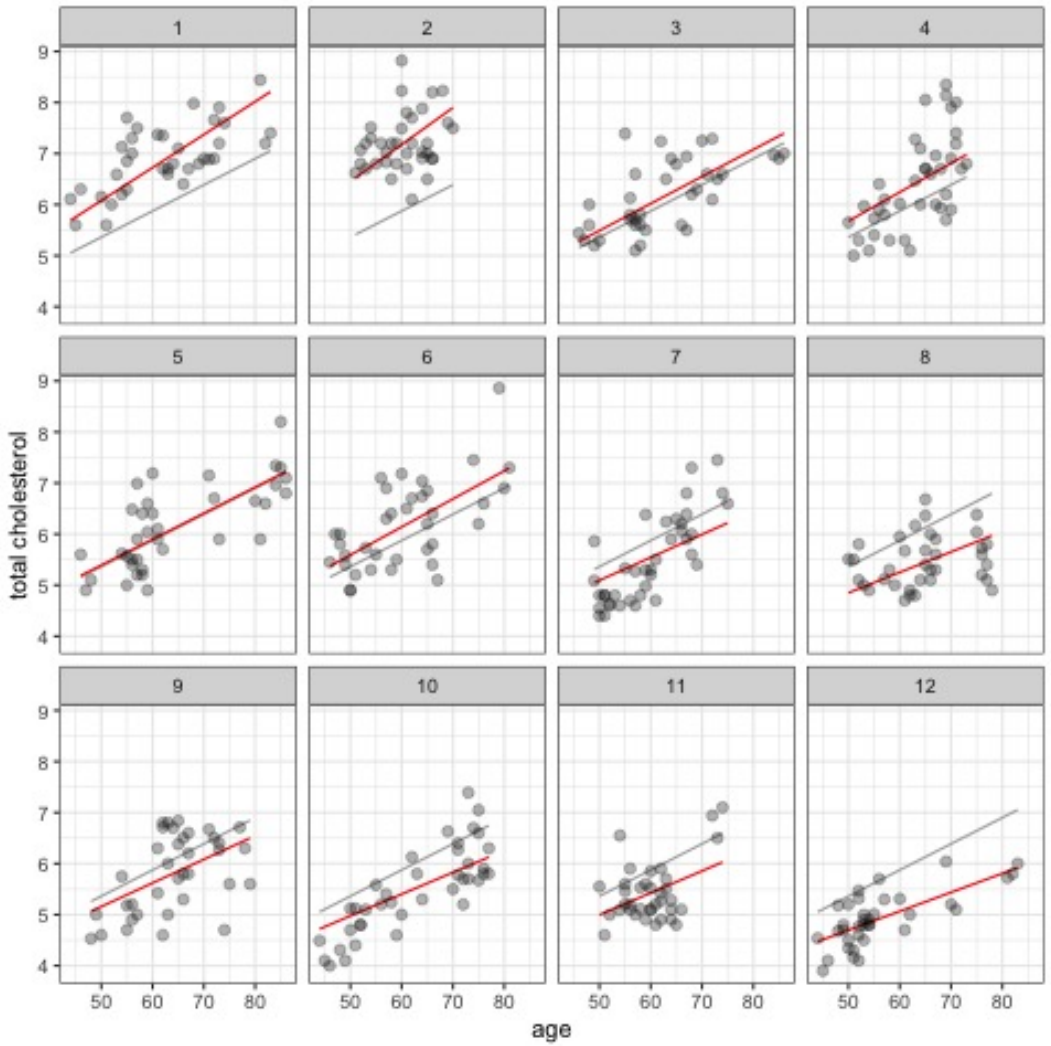
R: Group-specific slope in lm()

2. Group-specific slope

$$y_i = \alpha_0 + \beta_k x_i + \epsilon_i$$

- Add the group variable as an interaction with the predictor of interest.
- `lm(chol ~ age : as.factor(doctor), data=data.chol)`
- : only adds the interaction.

R: Group-specific slope in lm()



R: Fixed effects in lm()

How to specify formulas in the `lm()` function?

- Main formula: `y` $\sim$ `x`, where y is the outcome and x the predictor(s)
- Predictors can be added as:

$+$	main effect
$:$	interaction only
$\star$	main effect and intercept

Values:

- summary()
- coef()
- fitted()

Fixed effects: Disadvantages

- Fixed effects account for **any** unobserved group-specific confounders $\rightarrow$ Including both a group-specific intercept and slope is not identifiable.
  - When the intercept α_k is group-specific, then the slope is assumed to be the same for all groups.
  - When slope β_k is group-specific, then the intercept is assumed to be the same for all groups.
- If we add new groups to the dataset we may not consistently estimate α_k:
  - Consider α_1, the intercept for the first group.
  - When we add new groups, the slope may vary.
  - Changing slope will change the intercept, also α_1.
- Information on α_k or β_k comes only from observations in group k and we need to estimate one parameter per group.

Take away: Structured Data

- Most statistical methods are developed for independent and identically distributed (iid) data.
- But often in practice we observe structured data, where there is an intrinsic group structure.
- Grouping creates dependence: Observations within a group are likely to be more similar to each other than to observations from other groups.
- Ignoring the group structure can lead to over-confident results or even false positives.
- Analysing each group separately, we do not assume any shared mechanisms and need to fit a model on the samples within a group only.
- Aggregating and working only on the group-level drastically reduces the sample size k.